DOCUMENT RESUME

ED 041 053 TM 000 025

AUTHOR O'Reilly, Robert P.
TITLE State Education Department Leadership in Project and Regional Evaluation Systems.
SPONS AGENCY Massachusetts Univ., Amherst. School of Education.
PUB DATE Mar 70
NOTE 44p.; From Symposium "Designing Instructional Systems with Longitudinal Testing Using Item Sampling Techniques" (Annual meeting of the American Educational Research Association, Minneapolis, Minnesota, March, 1970)

EDRS PRICE EDRS Price MF-$0.25 HC Not Available from EDRS.
DESCRIPTORS Comparative Testing, Course Evaluation, Culturally Disadvantaged, Diagnostic Tests, Earth Science, Elementary School Mathematics, *Evaluation Techniques, General Science, *Individualized Instruction, *Program Evaluation, Research Projects, Slow Learners, Speech Handicapped, Speech Improvement, *Student Evaluation, Test Construction, *Testing
IDENTIFIERS *Comprehensive Achievement Monitoring (CAM), New York State Education Department, Pupil Evaluation Program (PEP)

ABSTRACT

The New York State Education Department's state-wide testing program using the Pupil Evaluation Program (PEP) is shown to be generally inadequate for judging program or school effectiveness, or for making decisions which would allow improvements in program effectiveness. The Comprehensive Achievement Monitoring (CAM) system can, however, be adapted to meet the information needs at different levels. Several experimental programs in which CAM models have been developed and implemented are described. Included are the use of CAM in research and development, demonstration projects in which CAM is utilized for course revision in convention classroom instruction, CAM combinations with other evaluative techniques for the purpose of evaluation of the work of individual students, and the use of CAM and small computers to evaluate student progress and assign students to instructional treatments. Finally, the support activities of the State Education Department to local school districts in the areas of program evaluation and management of the instructional process are outlined. (DG)

STATE EDUCATION DEPARTMENT LEADERSHIP IN PROJECT AND REGIONAL EVALUATION SYSTEMS*

Robert P. O'Reilly
Chief, Bureau of School and Cultural Research
Division of Research
New York State Education Department

*Paper prepared for a symposium, Designing Instructional Systems with Longitudinal Testing Using Item Sampling Techniques; Annual Meeting of the American Educational Research Association, March, 1970.

A recent paper by Campbell (1969) analyzed the political and methodological issues involved in efforts to evaluate large scale social and educational reforms. Campbell pointed to an increasing concern with effective evaluation procedures on the part of the political leadership and the program administrator, but concluded that most ameliorative programs end with no basis for an interpretable evaluation. There is evidence that the political and educational leadership in New York State has become increasingly concerned with developing the capability for hard-headed evaluations of educational programs supported under Federal or state funds. However, the stated objectives of program evaluations in New York State extend beyond the capability for simply judging program effectiveness. An analysis of the guidelines for project evaluation in the Urban Education Program (a special program for the improvement of instruction for educationally and socially disadvantaged children; 1969-70 appropriation of $52 million) shows evaluation procedures are to yield both effectiveness measures and information useful in determining strengths, weaknesses, or necessary modifications in individual projects. The New York State Pupil Evaluation Program (PEP), a state-wide testing program designed to evaluate Title I programs (1969-70 appropriation-$113 million) and assess achievement in all schools in the state assumes the capability to judge program effectiveness and to generate information useful in revising special projects and the school program (NYSED, 1967).

The objectives of approaches to evaluation in New York State education clearly reflect a concern with building a capability for making useful decisions about programs designed to ameliorate some of the more intransigent problems in contemporary schooling. This article examines the assumed decision making capabilities of the Education Department's

state-wide testing program (PEP) and other more typical evaluation procedures. The data collection procedures of the PEP and other approaches are then shown as generally inadequate for judging program or school effectiveness, or for making timely decisions which would allow administrators at different levels to gradually improve program effectiveness.

The subsequent discussion outlines the major features of the Comprehensive Achievement Monitoring (CAM) system, and shows how this evaluation procedure can be adapted to meet the information needs of program administrators and school personnel at different levels. There then follows a description of experimental programs in which the CAM model has been implemented to evaluate and modify programs, and manage the process of instruction in a number of ways. The examples presented include: (1) the use of CAM in R and D efforts; (2) demonstration projects in which CAM is used for course revision in conventional classroom instruction; (3) CAM combined with other evaluation approaches for the purpose of program revision and evaluation of the performance of individuals; and (4) extensions of CAM and other approaches wherein small computers are used in the schools to evaluate the instructional process and assign students or student groups to instructional treatments. Finally, this paper outlines the supporting activities conducted by the State Education Department (SED) for the purpose of assisting school districts and larger units in establishing more effective procedures for program evaluation and the management of the instructional process.

Evaluation and Decision Making Capabilities of the PEP and Other Approaches

Since 1965, New York has been involved in a landmark State-wide testing program called the Pupil Evaluation Program (PEP). Each fall,

all public and nonpublic school pupils in grades 1, 3, 6 and 9 have received certain standardized tests: a readiness test for grade 1 and tests in reading and arithmetic for grades 3, 6 and 9 (NYSED, 1968). The current testing program (NYSED, 1970) apparently does not include the readiness test, and there is some difficulty with the ninth grade tests which reportedly do not discriminate in the average and above-average achievement ranges (NYSED, 1968).

The annual PEP data are presumed to provide the basis for:

1. The identification of educational needs (NYSED, 1968), specifically schools in need of special supplementary funds (e.g., Title I) due to the presence of educationally and economically disadvantaged student populations.
2. The measurement of educational change, particularly in relation to the presumed impact of Title I programs (NYSED, 1970).
3. Making decisions relating to the improvement of instruction, budget making, supervision, allocation of personnel and the determination of educational quality (NYSED, 1970).

The basis for these identification and decision making capacities at the local level consists of raw score medians, percentile ranks and achievement levels (normalized stanines) calculated annually by school personnel. Staff for the PEP complete their own analyses of test results in which year to year comparisons are made in relation to the proportion of pupils "below minimum competence." The competence notion is defined as the 23rd percentile on all tests and reportedly has some

relation to the concept of competence, although no data have ever been offered on this point (NYSED, 1967).

Current analyses of PEP data, relating to the identification of the disadvantaged school and the potential impact of special aid therein, usually involve the examination of changes in the proportion of pupils below the minimum competence level (NYSED, 1967, 1968). The units of analysis are yet very gross, consisting of subject area (based on score totals for reading and arithmetic), grade level, and subdivisions of the population tested, such as public versus nonpublic students and urban versus rural students. Shifts in the proportions of pupils below the 23rd percentile, from year to year, are taken as evidence of improvement or regression, and as evidence of program effectiveness (NYSED, 1967_a).[1] Data for three years of the PEP generally show slight changes (both negative and positive) in state-wide estimates of proportions of pupils below minimum competence, and a tendency for these changes to be larger in smaller sampling units. Overall, positive and negative changes tend to cancel each other out when examined across grade levels.

With this brief description of the PEP testing program as a background we shall turn to an examination of its apparent capacity for decision making in the three areas previously designated.

Identifying Educational Needs

The PEP procedure provides a uniform standard for making relative school to school comparisons in two important areas of educational achievement. The relative needs of schools and districts may be

[1]Two subtest scores are available in reading, and three subtest scores can be derived for arithmetic. Norms are available for both total and subtest scores, but the latter have apparently not been used in the PEP analyses.

determined and special aid appropriately apportioned according to need. The identification of such needs, however, can be accomplished more efficiently through the use of sampling procedures. For example, the application of sampling procedures on test items and on students within individual schools would improve efficiency through reductions in testing time and costs for data analysis, paper, printing, and personnel. The use of systematic sampling procedures would still allow the required accuracy in determining estimates of proportions of individual students below minimum competence in individual schools, school achievement medians, and so on. Moreover, it is probable that the comprehensiveness of the testing could be increased (i.e., include more subjects, subtests, and grade levels), with the costs of testing remaining at or below current levels.

One objection to the use of sampling procedures for the identification problem would proceed from the fact that data on individuals would no longer be available for use in the schools. However, it is doubtful that this would be a serious problem, since the PEP data very probably represent a duplication of effort in the great majority of schools. The typical school has a testing program of its own, based on the use of standardized tests, which ordinarily result in information on achievement which is more comprehensive than that available from the PEP. Thus school personnel typically have the capability for the identification of individual achievement deficits, and can judge relative positions of individuals, classes and larger units in relation to national norms.

Educational Change and Program Effectiveness

An adequate evaluation of the efficacy of the PEP procedure for judging the effectiveness of special programs must consider the information

needs of the special audiences concerned with program evaluation. For large scale programs, such as Title I and the New York State Urban Education Program, important users include the political leadership, educational policy making boards and commissioners of education. Decisions at this level relate to annual appropriations for programs and the need to know of the effects of programs recommended by policy making boards. Another level would include State program administrators, who would require the same information as those at the uppermost level, but who may also require information which may allow them to make annual adjustments in the program. At a still lower level, school administrators and local program directors would appear to require information which would allow them to make annual adjustments in their own special programs, and/or select effective programs suitable to local educational needs. The information needs of these different audiences might be expressed as proceeding from the following determinations, among others:

1. Determining the general effects of large scale programs in broad areas such as reading and arithmetic.
2. Determining the relative effectiveness of different programs.
3. Determining the relative effectiveness of different program components (e.g., motivational vs. instructional components).
4. Determining the effects of different programs for different student groups (e.g., students grouped by ability, socioeconomic status, ethnic status).

The basis for these determinations consists of a series of annual data points which are used to make year-to-year comparisons. The more

serious defects of the PEP procedure may be made evident when the data are portrayed in a time series design, as shown in Figures 1 and 2. These illustrations show both a longitudinal analysis (Figure 2) and an analysis which corresponds more directly with the PEP approach (Figure 1). Both figures extend before and after the Title I intervention and show more data points than are available.

Figures 1 and 2 indicate one of the most useful features of the time series design--the sampling of instability. Performance is shown as a series of "ups" and "downs", with an overall trend for achievement to fall off, and then level out after the Title I intervention. There is the suggestion that performance is beginning to rise near the end of the series, but one could not be sure of this until more data points become available. The sources of instability seen in the times series design may include such factors as uneven effects of the treatment, differences between the population sampled at each data point, and regression effects. Shifts or changes in the trend of the data points may be evidence of treatment effects or other coordinated events.

When viewed in the context of the time series design, it may be seen that decisions about program effectiveness using PEP data are based on the unstable character of a series of annual data points. For example, the right-hand side of Figure 1 suggests no effect for a two-year period, then an effect, and then a decrement. A possibly more accurate portrayal, judging from the data now available, would be a series of slight negative and positive changes in values, resulting in alternate judgements of "effect" and "no effect." This is a "can't see the forest for trees" kind of analysis, yielding a comparable decision making capability.

Obviously, an appropriate procedure for judging program effects from

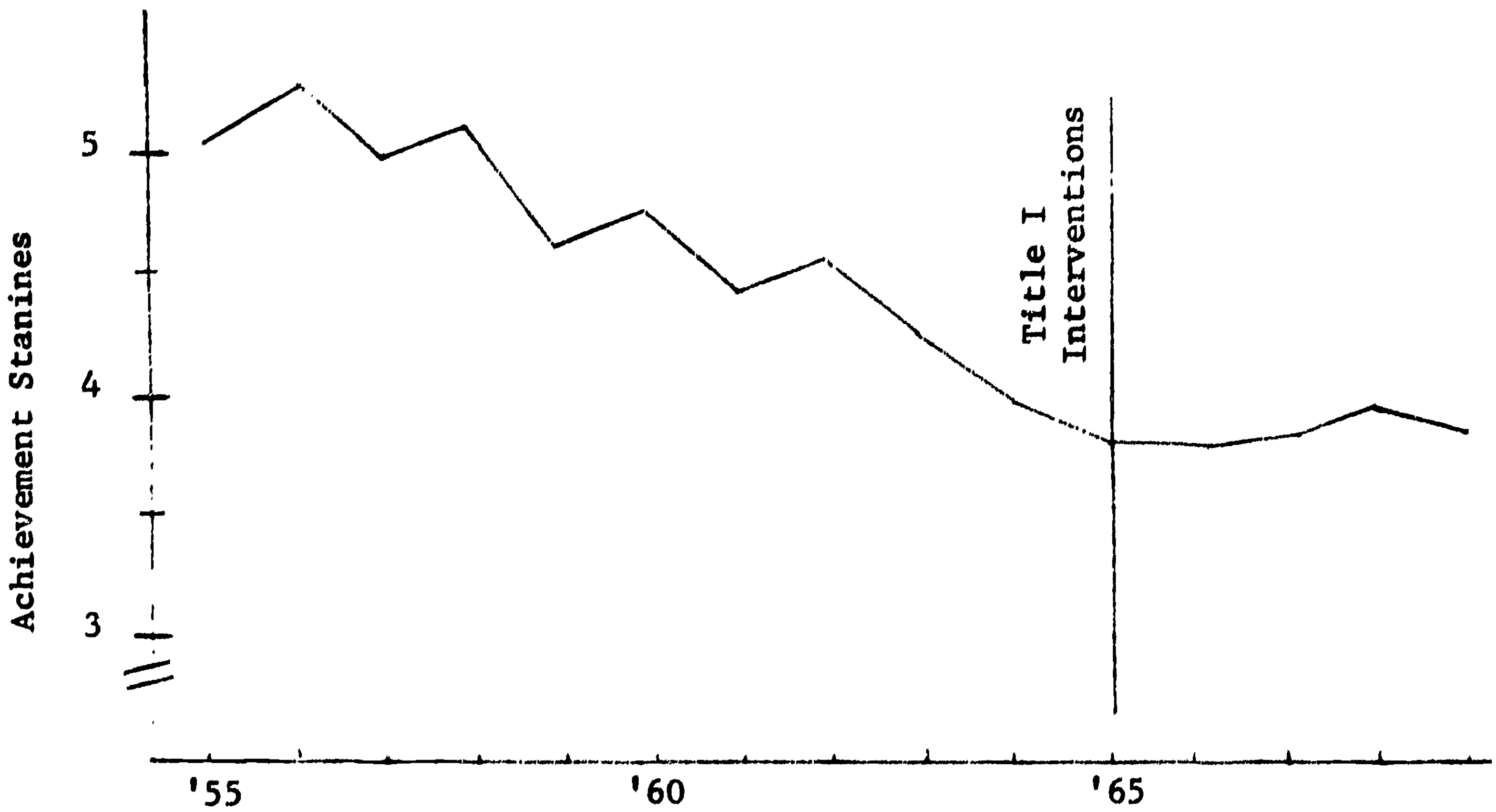

Figure 1: Average Reading Achievement Scores for Successive Third Grade Groups in Title I Schools (Imaginary Data)

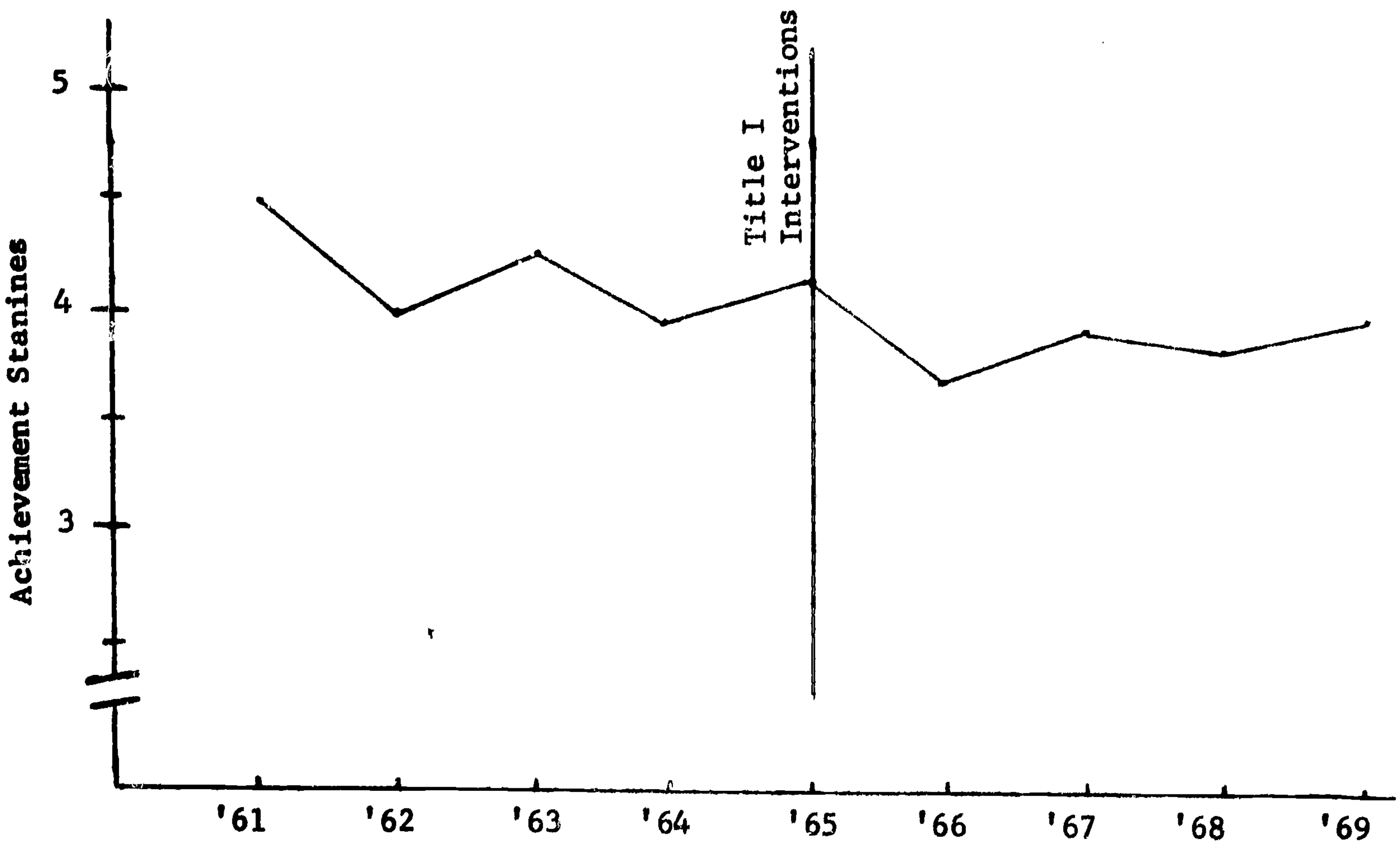

Figure 2: Average Reading Achievement Scores for Title I Pupils Followed for Successive Years in School (Imaginary Data)

the PEP would involve determination of trends or shifts within the perturbations of the time series. One sees here another serious problem relating to decision making capabilities: several data points are required in order for a trend to make itself evident. In systems like the PEP, this places the first potentially reliable decision point some years hence. Our analysis to this point thus indicates two major defects with procedures like the PEP: (1) the program manager appears to be placed in a position wherein his judgement about program effects represent the normal instabilities present in the data; and (2) the system is highly unreactive to the needs of the decision maker.

A further but perhaps less telling criticism of the PEP approach concerns its capability for determining the effectiveness of different programs and program components. To make such determinations, not only does one need valid procedures for indicating educational change, but the indicators must also be reactive to the particular objectives of the programs under evaluation. The latter requirement is not very likely met with typical standardized tests, such as those used in the PEP. Moreover, no testing program can meet this requirement as long as the testing concept is limited to the use of a single test or a small number of tests with necessarily limited numbers of items.

These problems with the determination of program effectiveness within the PEP procedure are not ordinarily circumvented by the addition of a supplementary but more specific evaluation program carried out by the local program director. The local evaluator tends to use the pretest-posttest design, perhaps supplemented by a nonequivalent comparison group. Aside from the fact that there are more threats to validity associated with this design than with the time series (Campbell, 1969), the problem

of sample attrition frequently becomes a further threat to validity, particularly where the program treats an urban disadvantaged population. The design may then reduce to a posttest only for the students who happen to be in the school at testing time at the end of the school year. Thus, in some urban districts, only a small portion of the treatment population may be available for both the pretest and posttest; the remaining posttest-only group would have varying degrees of exposure to the treatment.[2]

Other deficiencies of the typical pretest-posttest design result in severe limitations on the utility of the information available from testing. By using a criterion more specific to the objectives of the program, possibilities of detecting treatment effects may be enhanced. However, the limited sample of relevant behaviors available from two testing sessions is not ordinarily adequate for analysis of the effects of the components of the treatment. The lengthy time interval between testings may also result in a loss of meaningful information useful in improving the effectiveness of the program. For example, achievement of some important objectives may be transitory, perhaps due to interference from other events or to insufficient practice. Such effects constitute important knowledge about educational treatments, but they will generally not be evident in evaluation designs with eight to ten months between testings.

Improvement of Instruction

The third group of assumed capabilities in the PEP includes the generation of data useful for making decisions relating to the improve-

[2]The writer once selected a random sample of primary students, stratified by grade, age and sex in a pretest-posttest design in a large urban district. On the day recommended for the pretest, the examiners arrived to find only 25 percent of the sample available for the test. Experience suggests that attrition is frequently a serious threat to the usual approaches to study design in deteriorated urban schools.

ment of instruction. The current data gathering procedure would allow district staff to reach the conclusion that they should do something about their reading or mathematics programs. But, this type of judgement is usually possible with the data that schools routinely collect in their own testing programs. In any event, knowing that half the students in a particular district are below the 23rd percentile has little utility, when the questions of real interest to the school administrator and teacher are basically the same as those of interest to the legislator and program director. However, major differences between the information needs of school staff and those at higher levels would exist in relation to the degree of reactivity of the system, and the comprehensiveness and specificity of the information available from the system. For example, at one level of specificity, a teacher might want to make yearly judgements about whether certain components of her course or program are in need of adjustment (e.g., Are certain objectives being unnecessarily pursued? Are students failing to learn or demonstrating only temporary achievement of certain objectives?). Teachers will probably also want to know whether student groups with different ability levels require different instructional materials or approaches. Answers to similar questions, but referring perhaps to larger student groups and larger program components, would also seem to be of interest to the school administrator and program director. On an even higher level of reactivity, teachers and other school personnel might wish to make decisions relating to courses and student groups at intervals during the school year.

When it comes to making basic decisions about the instructional process, it is readily apparent that procedures like the PEP are outrun by even the simplest information requirements of teachers and school

administrators. The system does not result in enough data points or sufficient coverage of objectives to provide data useful for annual decisions. If we increase the data points per year to two, secure a stable student populations, increase testing time and thus improve coverage of objectives, and initiate carefully conceived controls, we then come closer to obtaining an improved basis for annual decisions relating to courses and student groups. However, even with these improvements, decisions on course revision may not be particularly reliable until two, three, or four years have passed. Decisions relating to student groups may be even more problematical if the characteristics of the student population tend to change in a particular direction every three or four years.

Some Further Considerations on Program Evaluation

Though the previous discussion has shown a number of serious inadequacies with the PEP, this should in no way reflect on the adequacy of the time series design for evaluating program effectiveness. Before processing to our description of CAM, which is based on a time series design, it should prove instructive to briefly consider the methodological adequacy and practical feasibility of the time series design in the educational setting.

As noted previously, the time series design consists of a series of equivalent observations (0), with a point of intervention (X):

0000X0000

Due to the fact that the times series has several pre- and posttests, all but one of several factors of internal validity are not considered a serious threat (Campbell, 1969). It is thus one of the more desirable

quasi-experimental designs, and is preferred to such designs as the one-group pretest-posttest design and other pretest-posttest designs involving separate samples. The one serious threat to internal validity, history, may be controlled through procedures which are practically feasible in the school setting (e.g., use of lagged control groups and carefully kept logs of events potentially affecting the treatment measure). Curve fitting or generating function procedures may be used to determine whether a statistically significant shift is present in the time series (Gottman, McFall, and Barnett, 1969).

The time series design, when combined with efficient data gathering procedures, is generally feasible in the context of large scale program evaluation, as will become more evident in the next section of this report. However, a few comments at this point will show some of its more useful features. The availability of several data points per year, for example, places the program administrator in a less ambiguous position regarding effectiveness judgements. The fluctuations of the time series, when it extends over a considerable length of time, may be used as an important source of hypotheses regarding the overall effects of a large program and the relative effects of individual programs. The effects of planned changes in a program may also be observed as the time series proceeds. Finally, through the use of sampling techniques, the time series provides the basis for methodologically adequate evaluation systems which would relate logically to the length of decision making intervals and the fineness of the decisions required by users at different levels.

New Approaches to Program Evaluation and Instructional Management

In recognition of the foregoing problems with the usual approaches to program evaluation and management of the improvement of the instructional process, the New York State Education Department has been experimenting with a variety of approaches derived from the CAM model, developed by William Gorth now at the University of Massachusetts. These activities are being conducted with support from Experimental and Innovative Programs, a State funded program directed toward the improvement of instruction in a variety of subject areas. What follows here is a brief account of the details of the CAM model and its potential benefits to users at various levels; a description of experimental CAM applications in New York State, and a summary of State Education Department activities undertaken for the purpose of generalizing CAM and similar activities to a wide variety of educational settings.

The CAM Model

We have already considered the inadequacies of one statewide evaluation scheme as well as the inadequacies of the evaluation procedures typically used by the local program manager, from the points of view of judging program effectiveness and obtaining reliable information usable in program or course revision. A previous publication (O'Reilly, Schriber, Gorth and Wightman, 1969) analyzed the decision making capabilities of the evaluation procedures typically used by teachers and concluded that these procedures have virtually no capacity for making systematic decisions about the instructional process. Similarly this paper concludes that the typical school testing program (based on the use of standardized tests) had very limited decision making utility for

revision of the instructional process, for many of the same reasons discussed in relation to the PEP system. The result of the limitations of testing programs at different levels in the educational establishment is a morass of overlapping data, obtained at considerable expense in time and money, but which possesses little practical utility for the program manager or those engaged in the instructional process.

The basis inadequacies of contemporary evaluation procedures, and of experimental designs which can be practically implemented in the schools on a large scale, are generally circumvented by application of the CAM approach. The CAM model is based on two ideas: (1) a flexible time series design which can be varied to meet the financial limitations and information needs of the user; and (2) a procedure for sampling students and items which introduces economy into testing, while at the same time increasing the comprehensiveness of behavior samples available from each testing session.

The typical CAM monitor is constructed around the stated objectives of the course or program to be evaluated. A number of test intervals is decided upon, depending upon the information needs of the user. A pool of items or other performance criteria is then constructed, with perhaps 4 to 10 samples per objective. Through the technique of random stratified sampling, items are assigned to test forms, thus creating a number of theoretically parallel test forms called monitors. Students receive test forms in a random order, at fixed testing intervals, until all tests have been taken by individual students. Test forms are typically short, usually taking from 10 to 30 minutes of the student's time.

The testing procedure wherein all students eventually receive all test forms over the period of a program or course is designed for

installations where periodic feedback on the achievement of individual students is desired. However, sampling procedures can be applied to schools, classrooms, and students to generate more efficient testing procedures. For example, larger installations might test in every school in the unit of interest, but random samples of classrooms and students within classrooms would be selected for each monitor.

One result of the CAM procedure is that behaviors relevant to a treatment or course are sampled at each data point. The use of sampling techniques further allows the program director or teacher to sample at each testing a much wider variety of behaviors than is normally possible with the usual evaluation designs employing standardized tests. The resultant data for specific instructional objectives or program components may be arrayed in an extended time series design as shown in Figure 3.

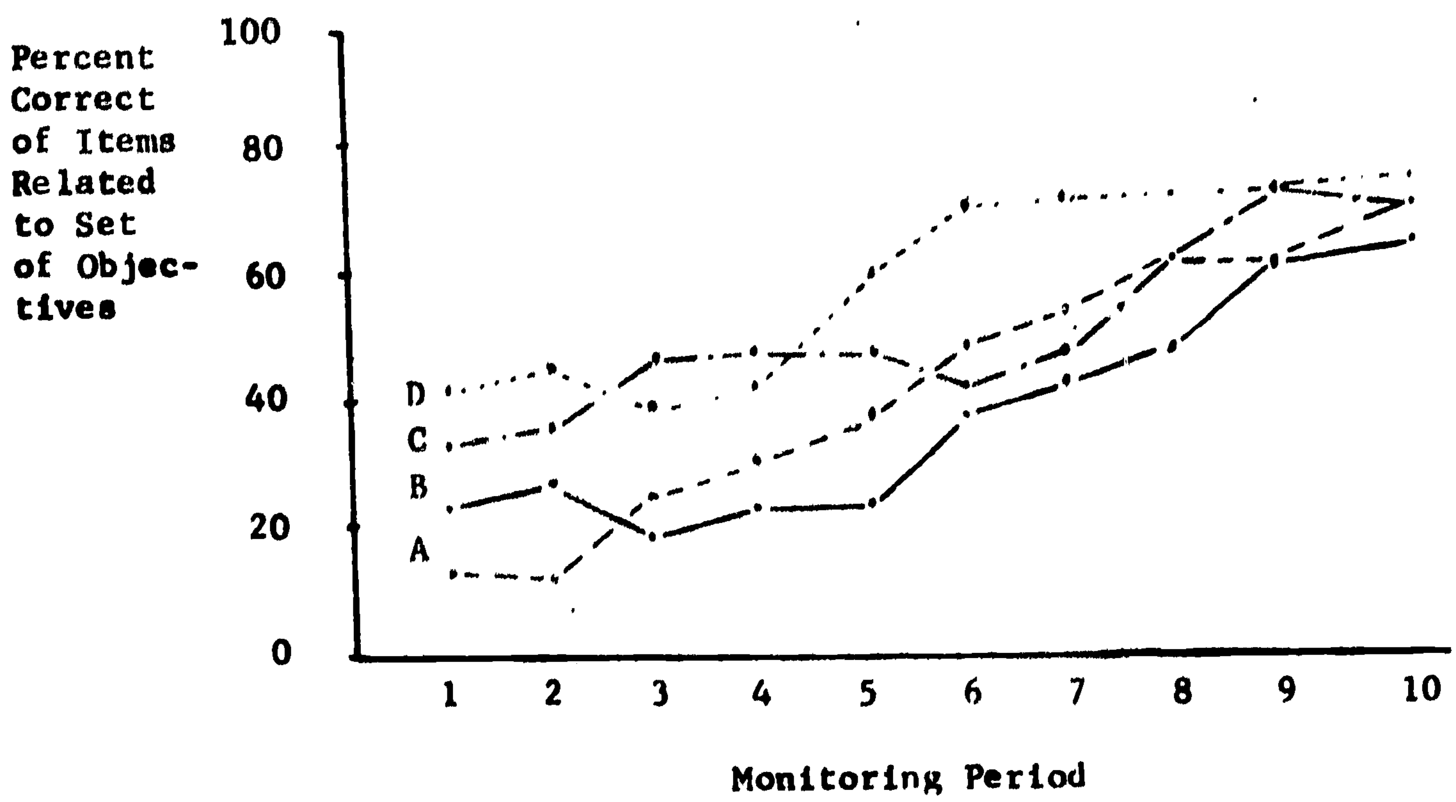

Figure 3: Illustration of CAM's capability to monitor performance on any selected set of objectives or any subject area units.

The point of intervention of an instructional component may be easily identified in relation to a particular data point or series of data points. Points of intervention which occur later in the series may thus form the basis for an interrupted time series design. The series of data points from a CAM monitor may also be arrayed in relation to a variety of meaningful subgroups as shown in Figure 4. The four series given in Figure 4, for example, illustrate the capability to follow the performance of any student group (or individual) across an entire course or treatment. The symbolic A, B, C, D (etc.) can be any

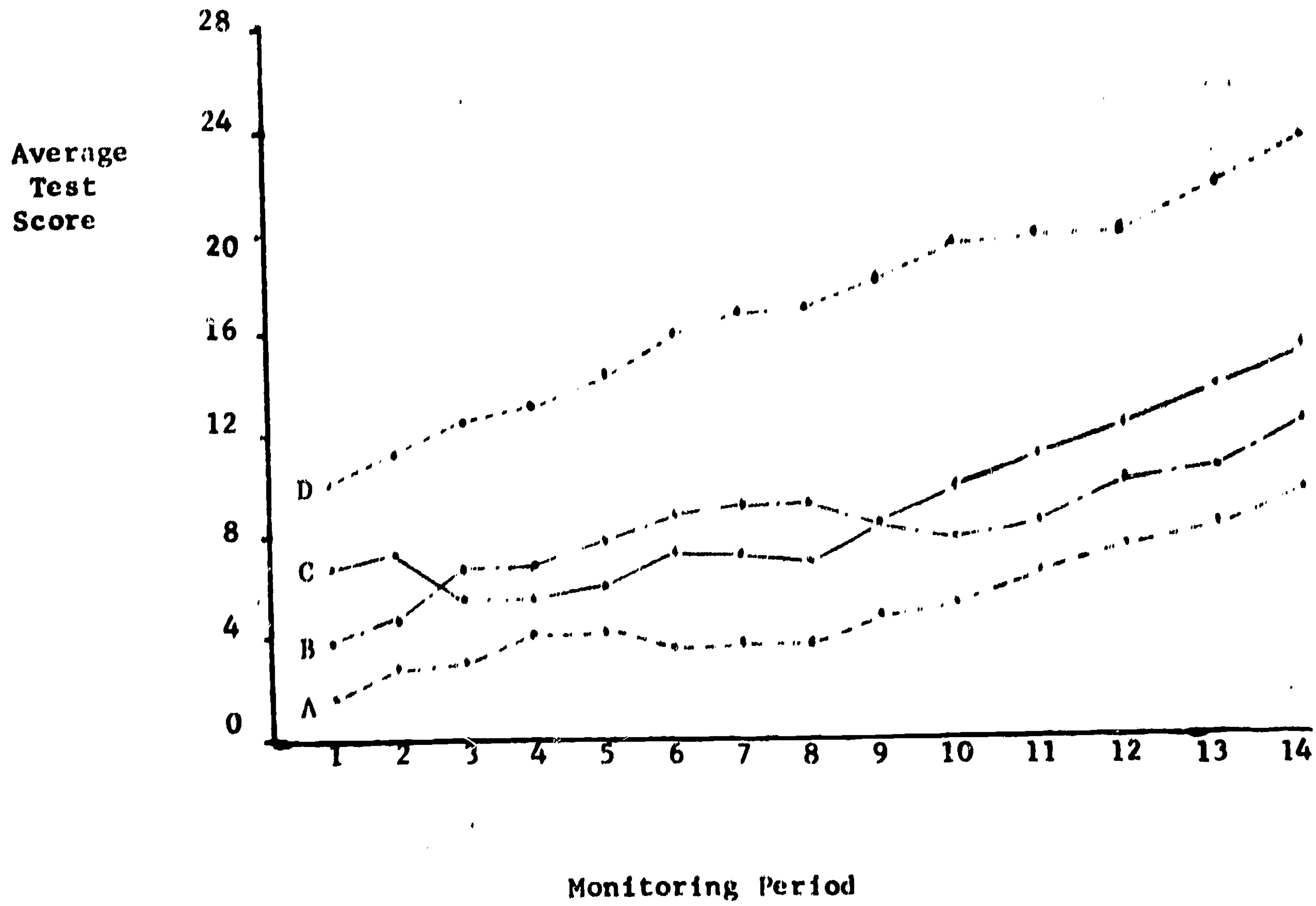

Figure 4: Graphic Representation of CAM's Ability to Follow and Examine Performance Progress of Groups or Subgroups Continuously

combination of students. Several possible levels of focus on performance would have the A, B, C and D symbolizing performance by:

1. Class or grade quartiles by ability (as determined at the start of the course).
2. Different classes taught by the same teacher.
3. Groups contained in one classroom, e.g., students grouped for instruction in a particular subject area or skill.
4. Groups being treated by various teaching approaches.
5. Individuals.
6. Experimental and control groups.

Some Advantages of CAM

The CAM model is fully operational in a number of settings in New York and other states. A set of computer programs is available to generate a variety of forms of data at low cost, including the achievement profiles shown in Figures 3 and 4, individual student reports, summary data for course units or program components, and item analyses. Depending upon the number of data points included in a monitoring system, and the comprehensiveness of the items across test forms, the data allow the following types of judgements in relation to points in time and specific events in a course or treatment:

1. Short- and long-term retention.
2. Pretest information on all objectives.
3. Increments in achievement over time.
4. Interactions between objectives or program components (i.e., teaching A affects achievement on B).

5. Subgroup achievement.

6. Group by treatment interactions.

With a sufficient number of data points in a course or treatment, the outstanding problems with procedures like the PEP and the pretest-posttest design are solved. For example, the problem of sample attrition becomes less of a threat since there are likely to be at least two or more data points available for each student in the treatment or course. Due to the large number of objectives which may be tested, and the typically high degree of specificity of the tests employed, the CAM procedure is likely to be much more reactive to the effects of particular events included in a complex treatment. The CAM procedure thus results in a potentially more effective evaluation of program effectiveness, relative to the procedures typically used, and also provides information useful in restructuring a treatment or course.

Use in Statewide Evaluations

The discussion has previously shown how systems like the PEP can be rendered more economical by the introduction of sampling procedures. By using sampling procedures on schools, programs and students, it is possible to generate monitoring procedures which can meet the information needs of users at one or several levels (Gorth, Dumont and Wightman, 1969). For example, the PEP procedure could be replaced by a monitoring system with four to six annual data points, probably without increasing total testing time or costs. The comprehensiveness of the evaluation procedure (i.e., number of objectives assessed) would be increased, and the design could also include the potential for stratifying by student group and type of program.

Monitoring procedures may be combined with other efforts at

information gathering. For example, the State Education Department's Information Center on Education (ICE) routinely collects information on instructional factors and teachers, and ethnic data on students in every school in the State. ICE staff are currently considering the collection of family background data on students, similar to the type of data collected by Coleman et al. (1966). The annual ICE data, when combined with achievement data from a statewide CAM, would create numerous possibilities for stratifying the student sample in relation to qualitative school factors, special programs, and individual differences, such as ethnic and socioeconomic status.

The CAM procedure can also be used to effectively monitor the progress of individual programs or schools, by shifting the unit of analysis and generating additional test items related to the more specific local objectives. If there is special interest in particular programs supported under State or Federal funds, package tests or sets of items specific to a program can be incorporated within the overall monitoring system.

A further advantage of CAM over traditional evaluation procedures resides in its feedback capability at various levels. A statewide CAM, such as is briefly alluded to here, can be used to generate profile data in several areas of achievement on an annual basis. The resultant data may be used at the state and district level to judge effectiveness and make adjustments in special programs and in the programs offered by the school. The potential effects of such adjustments may then be followed in successive years via the monitoring procedure.

This feedback feature of the CAM system can be extended downward to the individual teacher and student. Experience with applications

of CAM in courses suggests that the periodic feedback element is a strong force for motivating improvement in teachers and students. Periodic student feedback on specific course objectives may be particularly effective in the disadvantaged school, where current research shows that students typically lack a sense of control over events affecting them. Wilson's (1967) report suggests that this attitude results in part from a lack of success in school. One might reasonably suppose that sense of control is affected by the meaningfulness of feedback in achievement situations. The potentially more meaningful feedback inherent in the CAM data, as contrasted with report card grades and standardized test results, suggests that classroom applications of CAM would be a worthwhile experiment in the disadvantaged school.

Use of CAM in Project Evaluation and Development

Experimental Programs for 1970-71 include development activities designed to radically change some components of the educational process. We are currently including CAM type evaluations in all experimental programs, and have initiated a CAM approach on a trial basis for the 1969-70 and 1970-71 school years of our Levittown Laboratory Mathematics Program. We expect to initiate a CAM type evaluation in our developing Program Speech Improvement System (PSIS). A brief discussion of our CAM related activities as they are being initiated or considered in relation to these two programs will show how CAM is particularly effective in R and D efforts.

The PSIS program is concerned with the development of automated instructional procedures designed to correct articulation difficulties (incorrect production of specific speech sounds) in young children, an

area of concern which comprises roughly 80 to 90 percent of the speech therapist's instructional activity. Estimates based on analysis of data from the Basic Educational Data System (a SED information system) for the 1967-68 school year indicated that approximately 114,000 students received speech correction services in New York State schools at a cost of $31 million. To serve the additional 86,000 students in need of speech therapy services would have required an additional 790 speech therapists and a further expenditure of $23 million. Since there is little likelihood that instructional needs in this area will be met in the future using conventional approaches, it seemed particularly appropriate to consider the development of alternate instructional procedures which may extend the effectiveness of the conventional speech therapist.

To remedy the disparity between the availability of the speech and communication expert and the order of need for his services, an initial low budget program was begun at Ossining, New York, involving the development of an automated speech instruction program adaptable for use with personnel lacking specialized professional training. The investigators for this program developed and collated materials in a programmed form suitable for presentation through a variety of controlled audiovisual input-output devices. One of the most innovative features of the program is the economy of operation made possible through the use of auto-instructional techniques and the services of teacher aides.

During 1968-69, an experimental program prototype designed to correct articulation difficulties of children in grades 1 to 6 was partially completed and subjected to test. During 1969-70, additional program components are undergoing completion, including a complete system for evaluating student performance, introductory training

materials for training the child in responding to the automated sections of the program, a cost effectiveness analysis, and a complete delineation of the teacher aide and therapist roles. In addition, the complete program is being tested and refined in three experimental settings, with one of the settings involving urban minority group children.

During 1970-71, it is expected that the program refined over the previous year will be further refined in a field test in ten different school settings, including schools in large urban areas. A complete evaluation of program effectiveness, including a cost analysis, will become available at the conclusion of this field test year.

We expect to initiate a CAM-type evaluation in the PSIS field test in 1969-70. Previously, we have used the pretest-posttest, control group design, with the dependent variable being sound production. At this point, however, we hope to obtain more detailed information on the effectiveness of the program components. The size of the problem becomes evident when one notes that the program has four major training phases (gross auditory identification, fine auditory discrimination, sound production and sound stabilization) for each of 14 different sounds. Evaluation of the stabilization function, and the related phenomenon of spontaneous correction of speech difficulties, becomes particularly problematical in the usual pretest-posttest design. For example, many students experience the program for only a few short weeks. An adequate evaluation might require testing each treated pupil at the initiation and conclusion of instruction, and then at one or more followup testings. More than one followup testing would be desirable in order to obtain some rough determination of the point of possible breakdown in sound production. There are the additional problems in this particular project

of the expense of obtaining adequate measures of sound discrimination and speech production, and of evaluating other aspects of the program (e.g., the therapist and aide roles).

By fully utilizing the random elements of the CAM approach, we expect to achieve the major goals of evaluating program effectiveness and obtaining information relevant to revision of the program at the conclusion of the field test year. Mainly, what is required is to coordinate the collection of CAM monitors with the average amount of time students spend in a treatment. By using sampling procedures, it will be possible to monitor all program components at each testing time. The CAM procedure should result in the generation of ideal data on stabilization of sound production and spontaneous correction of faulty sound production. Evaluation plans also include keeping logs of all significant events in program implementation in all field test settings. These data may be particularly meaningful when the time comes for examination of the profile data. The combined results of the CAM and log data may further allow us to make potentially meaningful adjustments in the project as the field test year proceeds.

Another project, the Levittown Laboratory Processes in Mathematics Instruction, includes a CAM system for the dual purposes of program refinement in the final stages of development, and experimental installation of the CAM approach in a BOCES (Board of Cooperative Educational Services) facility. The BOCES facilities in New York State offer shared services to school districts in particular counties in New York, with such services including the capabilities of modern data processing centers.

The Levittown Mathematics Program consists of the use of calculator-assisted, individualized instruction to aid slow learners in mathematics

and employs CAM as an evaluation implement for nearly 600 of the 800 pupils using calculators. CAM was initiated into the program in the Fall of 1969. The CAM design and components were formulated and completed during the preceding summer by a team of five mathematics teachers and coordinators who have been part of the laboratory processes program from its inception. With the assistance of Department staff, two CAM systems were created. Each was constructed to be used over a two-grade span and consisted of 180 instructional objectives, an item bank of 700 items, 18 test forms of 36 items each to be administered bi-weekly, and data analysis done by computer. CAM I was designed for use in the fifth and sixth grades and CAM II for use in the seventh and eighth grades.

The creation of the CAMs required a concerted team effort of seven weeks duration to formulate the objectives and items and to construct the test forms. Department staff served in an active advisory capacity working directly with the team, with assistance extending from introducing the CAM process to the team through the construction of the test forms. Department staff also provided training and techniques in objective and item construction to the team members. The major effort involved the formulation of objectives and items which required a restructuring and resequencing of the curriculum. Topics were selected and general objectives formulated. From the general objectives, sequential sets of specific behavioral objectives were created. The specific objectives, in turn, were used as the basis for item construction. Other efforts involved the mechanical operation of compiling items to create test forms, the clerical task of preparing the test forms for reproduction, distribution of tests to teachers, and orientation of teachers to the CAM

process.

Efforts are being made to adapt the analyses of the CAM test data to the local computer facilities of the Nassau County BOCES. Until such adaptation can be completed, data are being analyzed through facilities at the University of Massachusetts. These facilities and other technical resources for CAM are available through William Gorth of the University of Massachusetts, originator of the CAM procedure.

Data available to teachers is in the form of achievement profiles, individual student reports, and item analyses. Achievement profiles delineate group and subgroup performance as the groups move through the course of study. Periodic determination of group progress and course effectiveness is made possible through a continuous feedback of data. Individual student reports are returned to the student after every testing and show him which objectives were represented on his test form and whether or not he correctly answered items related to each of the objectives. This feedback permits immediate followup in the form of individual student-teacher conferences to determine individual strengths and weaknesses and to prescribe specific instructional remedies when necessary. The item analyses consist of a year-end compendium of student performance on items related to three time phases: items encountered before, after, and at later retention periods in relation to presentation of the objective. Thus, the analyses not only afford a discrimination index of difficulty and suitability to student abilities for each item, but through the three phases they provide further indication of an item's effectiveness and a pattern of group mastery of each item and its related objective.

In addition to instituting CAM, the Levittown program also includes a modified experimental design to assess the effects of CAM versus

conventional evaluation techniques. Four treatments groups are involved: (1) students using both calculators and CAM, (2) students using CAM only, (3) students using calculators only, and (4) students with neither CAM nor calculators. The experimental design includes over 1,300 pupils in five school districts.

The data from the Levittown CAMs for 1969-70 will become the basis for program adjustment during the coming summer, resulting in further revision of the program, new manuals of instruction, and refinement of instructional activities. During the 1970-71 school year, a focus on program validation and refinement will continue. At the same time, plans will be implemented to make the evaluation component more reactive to the individual student. This will be accomplished by creating separate one-semester CAM monitors for each grade level treated. Key terminal performance objectives will be overlapped between semesters, but the overall result of the plan will be to pretest students on objectives for one semester at a time, as compared to the current procedure of periodically testing on all objectives over a two-year span. The plan also includes phasing out multiple choice items in favor of the constructed response format and further increasing item validity by accounting for computational and other irrelevant errors in test scoring.

Improving Instruction: CAM Applications

In addition to the Levittown installation, which serves both R and D and model demonstration program purposes, we have two additional CAM installations designed primarily to monitor the instructional process. The program at the Westchester BOCES No. 1 is a typical CAM system designed to yield information on the instructional process in conventional mathematics instruction at the fifth and sixth grade levels. The

program now affects about 850 students and will be extended to the reading area during 1970-71. We intend to limit the program in the BOCES to conventional bi-weekly monitors, but will extend the monitors to other grade levels and all key subject areas. Future plans include adaptation of the CAM software for use in the BOCES IBM 360-40 computer, with the BOCES eventually achieving the capability of offering CAM course monitors as a regular service to the schools it serves. Such services will include printing student report cards, items analyses, and achievement profiles. It should be mentioned that preparation and training for implementation of the Westchester BOCES installation had many positive effects on the current instructional program: (1) behavioral objectives were written for the first time in the elementary mathematics program, and (2) both students and teachers were able to use the instructional objectives as a guide in teaching and learning. Additional positive effects of the CAM approach should become apparent when the program is revised during the coming Summer.

Our second CAM installation, designed to monitor the instructional process, serves 300 sixth and seventh graders in the Ballston Spa Middle School. The subject area treated is again mathematics; there is a separate CAM for each grade level; and 16 monitor periods in each CAM. Initially, data for this program were analyzed by hand, but we shortly shifted to computer analyses.

The Ballston Spa program is totally individualized in all key subject areas, and for this reason tentative plans have been formulated to create a system for both monitoring the instructional program and prescribing the learning activities of individual students. The new CAM system for 1970-71 will thus fully utilize sampling techniques, will

include only 4 to 6 data points instead of 16, and will only require about 160 minutes of testing time, per student, for the school year. Student feedback from the CAM monitors will be deemphasized in favor of utilizing the data for making such determinations as strong and weak points of the course performance by student ability group, and retention of important instructional objectives.

In addition to the CAM group monitoring system, the project will be concerned with the creation of a system for placing the student at the appropriate point in the course of instructional objectives, and then determining the adequacy of his performance as he proceeds through the course. A tentative plan of the procedures for making instructional decisions about individual students is shown in Table 1.

The test schedule for individual decisions shown in Table 1 is based on a series of six-item subtests, one subtest for each of 100 instructional objectives. The objectives in this program are all complex terminal performance objectives and are arranged in a hierarchy. Students are initially given a pretest, containing 60 items, with three items for each objective in the first two (or more) of the instructional units. An instructional unit is a learning package which can be completed in a few days or weeks. If the student passes all items in the pretest, he is given the third unit test, and continues with the testing procedure until he misses two or more items in a subtest. If he misses two items in a given subtest, he may be given additional items in order to gain a more accurate determination of his placement. Students who miss more than two items in a given subtest have tentatively defined their starting point in the instructional hierarchy. To insure accurate placement, at the starting point, the student may be required to take one or more subtests which come

Table 1

Possible Test Schedules for Cam Monitors and Assignment of Students to Instructional Treatments

	Test Schedule Individual Decisions			CAM Monitors Group Based Data	
Instruction Schedule	No. Obj's	Items/Obj.	Items/Test	No. Items/Testing	No. Obj's/Test
Pretest (Start)	10	3	60	400	100
Unit I	4	6	24		
Unit II	6	6	36		
Unit III	4	6	24		
Unit IV	4	6	24		
Unit V	6	6	36		
Unit VI	1	6	6		
Unit N	2	6	12	400	100
Total	100	600		400	100
Notes	Number of Testings variable from course to course; test schedule not fixed.			Number of tests, forms, and items per form based on student sample of 20-30 for each item, each testing, test schedule fixed.	

earlier in the hierarchy (units one or two).

Once the student has been placed in the instructional hierarchy, he will take subtests as he completes a unit or portion of a unit. At each testing session, his test will be immediately scored and a decision made which will result in his receiving either remedial activities or the next instructional unit. The criterion for moving to the next unit is a

minimum of five out of six items correct on each subtest. Performance below this criterion results in the student receiving an individual diagnosis and then a prescription. Diagnostic procedures will repeat the assessment of the relevant terminal performance objectives and of enabling objectives as far as is necessary to determine the cause of inadequate performance.

The creation of the subtests included in the individual testing system and their arrangement in a hierarchy will be based as far as possible on the CAM data from 1969-70, which include pretest and posttest difficulty levels. The correctness of decisions based on subtest performance will be enhanced by using the constructed response format for items and by scoring for irrelevant errors.

The validity of decisions to release individuals from treatments may be investigated by varying the decision making basis (e.g., one item versus two items missed on an objective) and then contrasting the decisions with retention data available from the CAM testing. The responsiveness of items and subtests to treatments may be determined by rotating items from the CAM system to the six-item subtests and then back again, on a regular basis. This procedure would eventually yield pretest and posttest data on every item included in the pool. Additional analyses would also be required to investigate subtest homogeneity.

Subtests in the individual testing system are to be printed on separate cards, coded by objective and kept in an achievement monitoring center. The monitoring center will be manned constantly to maintain test security. The routine mechanics of testing will be handled by a teacher aide. Diagnosis and prescription will be essentially automatic for students who meet the five-sixths criterion. The teachers' role will focus basically

on the tasks of diagnosing the causes of inadequate performance, assigning students to remedial treatments, and individual tutoring. A standard record will be kept of the procedures used in each diagnostic session in the hope that ways to improve the procedures can be discovered.

Computer Based Systems: Current and Planned

During 1969-70, we began our first experimental attempt at using a small computer to analyze and print student achievement data on a day-to-day basis. The initial operation is being conducted at the Greece Central Schools, under the direction of Dan Heisey of the University of New Hampshire.

After a careful analysis of hardware and time sharing costs, Digital's PDP-12 computer (4096 words, 12-bit core memory) was secured on a rental basis for one year. Cost for the machine and a heavy-duty teletype unit with card input capabilities on a three-year, lease purchase contract is $13,110 per year. The service contract and insurance raise the cost to $15,830 per year. A high speed printer may be added for approximately $4,000 per year on a lease purchase basis. Cost of the system thus ranges from about $48,000 to $60,000, depending on the components desired. The unit will eventually be used to process achievement data for approximately 8,000 students and will produce savings relative to a time sharing facility or the batch processing mode now used for our other CAM installations.

The computer is scheduled for installation in the Greece schools by November 1970. In the meantime, the necessary software are being developed and run on a trial basis on the PDP-12. One school is now receiving regular output which is being run on a time sharing facility. The time sharing facility is being used to generate more efficient versions of

the programs, which are then adjusted for use in the PDP-12. The content and format of the output are being checked with the teachers as the programming proceeds.

The Greece assessment program now covers mathematics in grades K through six. Like the Ballston Spa program, objectives are arranged in a hierarchy, although this assessment procedure is designed to monitor only basic skills. The number of objectives ranges from 7 at the kindergarten level to more than 50 at the sixth grade level. The format of the tests resembles that used at Ballston Spa in that each objective is assessed by a six item subtest. However, items are grouped into 20 item tests and include an overlapping CAM which is designed to obtain pretest and long- and short-term retention data on all objectives as the course proceeds. Output thus gives teachers the capability of determining students' learning needs as well as the overall effectiveness of the course.

The programming is also being designed to partition students into instructional groups, which are then assigned to the teacher team for more detailed diagnoses and treatment. As now formulated, pretest data and test records from the previous year will be used to form instructional groups defined in relation to terminal performance objectives as shown in Figure 5.

As shown in Figure 5, student groups are taught one or two objectives by a teacher team member. At the conclusion of instruction, students receive an achievement monitor, data is processed within 24 hours, and new groups are formed. Both students and teachers receive a test record indicating item performance by objective. The procedure as shown in Figure 5 requires teacher aides, a teacher team whose members are free-floating, a series of diagnostic tests designed to assess the

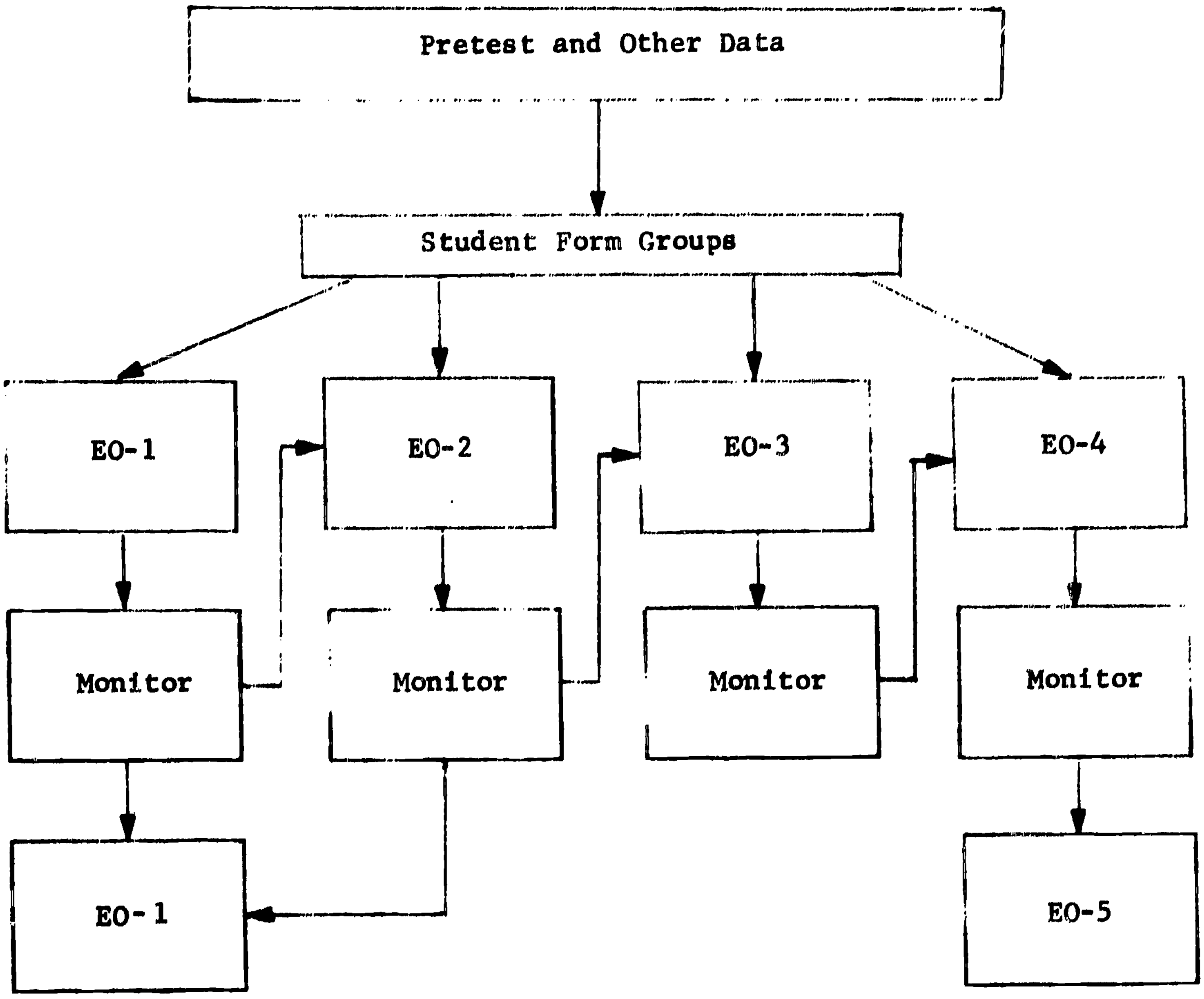

Figure 5: Partioning Students
in the Greece Assessment Program

details of inadequate performance on each objective and a greater variety of materials than is usually used in conventional mathematics instruction at the elementary level.

We do not yet know how smoothly this partitioning scheme is going to work. Most of the work being done on the program at present is concerned with developing a flexible set of programs which can be used on the PDP-12 and will produce basically the same type of output which can now be accomplished only on large computers. We expect this program to demonstrate cost effective potential in two ways. First, the current cost of CAM monitors, using the batch processing mode, is \$2 to \$3 per student, per year. The PDP-12 should bring this cost closer to \$1 per student, when averaged over a period of five years. Secondly, we expect that teachers will eventually perform more of a management than a teaching function. By collecting appropriate instructional materials and using brief, curriculum embedded tests, a linear path can be arranged for instructional groups at the upper primary and intermediate grade levels. By using aides to distribute materials and perform routine daily monitoring functions, the teacher should be able to effectively handle more students. The costs of the monitoring system should thus be offset by staff savings and the effectiveness of instruction should theoretically be improved.

Our latest effort to develop useful systems for managing the instructional process will be initiated in the Summer of 1970 in the Jamesville-DeWitt Individualized Science Program. The Jamesville-DeWitt Public Schools have developed a programmed and semi-automated system for teaching general science, earth science, and advanced general science. The completed system allows the advanced scien e student to complete as

much as three years of science instruction in a single school year, through the use of programmed instruction devices which allow the student to move at his own rate. Similarly, the slower student is enabled to complete basic ninth grade science instruction within the normally allotted time. The system is totally individualized, 80 percent automated in its instructional aspects, and results in savings by expanding the pupil-teacher ratio.

Recent research indicates that no significant differences in student achievement results when instruction through independent study techniques is used instead of conventional classroom methods. However, when independent study is utilized to implement individually prescribed instruction:

1. The increase in achievement is significant at the .05 level of confidence or beyond.
2. Required instruction time is often reduced by more than half.
3. Student efficiency is greatly increased.

The model program of individually prescribed instruction developed by the Jamesville-DeWitt Public Schools has been designed to realize these advantages of a combined independent study and individually prescribed instructional program. Courses in which the model is to be implemented are organized into "modules" which consist of a series of events arranged in a hierarchy of behavioral objectives leading to mastery of a single concept or group of related concepts. Each of these instructional events constitutes a "learning activity package" or LAP. Each LAP is designed according to the flow diagram shown in Figure 6.

The module pretest is used as a primary diagnostic tool along with

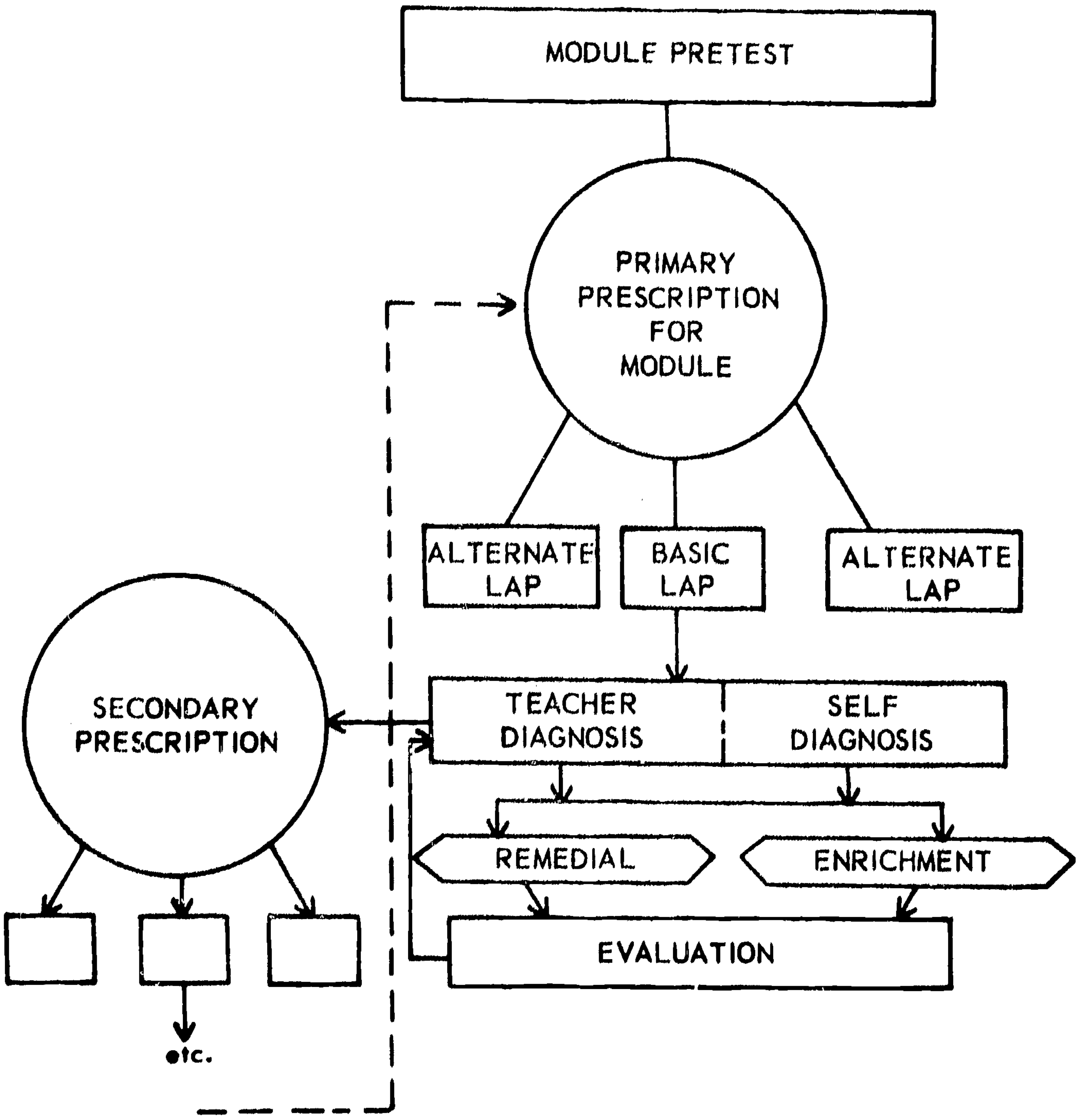

Figure 6: Flow Diagram of the Jamesville-DeWitt Learning Activity Package.

the results of other social, intellectual, interest, and achievement evaluations of each pupil to select the proper type of LAP and the hierarchy of LAPs in the module. Each LAP is designed around as small a number of behavioral objectives as possible. Secondary diagnosis is currently built into each LAP to provide both self testing by the student and criteria for a higher level of diagnosis by the teacher.

The LAP approach to teaching results in an even more flexible and efficient model when related courses are combined around a set of core objectives, as shown for the ninth grade science program in Figure 7. Core objectives which are common to general science, earth science and advanced (honors) general science form the basic objectives of the model.

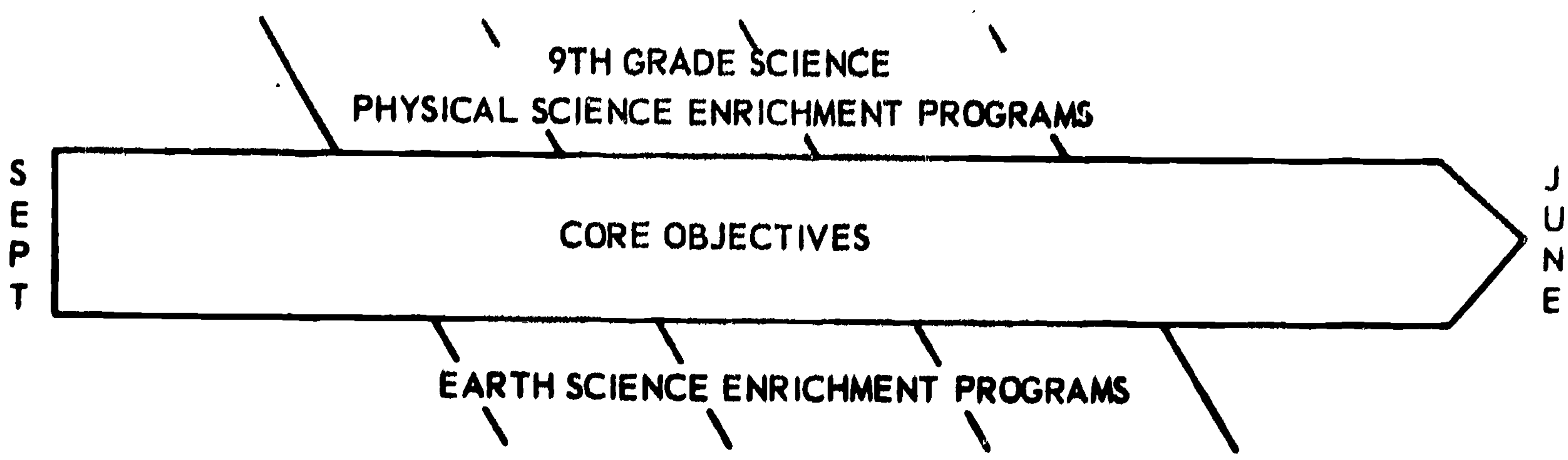

Figure 7: Arrangement of Core and Advanced Objectives in the Jamesville-DeWitt Science Program

The Jamesville-DeWitt science program and other courses built around the LAP approach have demonstrated cost savings in professional personnel. However, each program still requires a substantial investment in time and funds to train teachers to adequately perform the testing, diagnostic and prescription functions. As staff members move or are replaced, training

in these functions becomes a continual problem. Variation in teacher capabilities in diagnosis and prescription also introduces a looseness into the system which is difficult to control. For these reasons, we intend to bring the basic functions of student placement in the LAP, diagnosis and prescription, under control of a small computer such as the PDP-12.

Supporting Activities

Obviously one of the problems with implementing CAM in the school setting is the amount of work which must be done in such areas as curriculum analysis, writing behavioral objectives, writing and field testing items, and preparing and printing tests. To ease problems in this area, we have begun to collect a bank of objectives and items from our own projects and from other installations around the country. As our programs continue to develop, we shall be able to develop a comprehensive objective and item bank complete with relevant item analysis data. This information can then be provided upon request to any district interested in starting a CAM.

A second major problem with CAM concerns the training which must be offered to school personnel in order to get the system underway. To this end, we have been distributing training materials which present the technical details for creating the monitors, and additional materials which serve as a guide to the preparation of behavioral objectives and test items. Consultant services, however, are still required to initiate the construction of CAM systems and to monitor progress through the first year. Our new CAM manual, being developed in cooperation with William Gorth, will be out around June 1970, and should obviate some of the need for consultant services. The new manual will be in more of a cookbook

style than our previous efforts, and will contain more detailed material relating to the interpretation of data, preparation of objectives, and making decisions about course and program revisions.

Another area of concern is the availability of software for computer analysis of the CAM data. To that end we have been working on the adaptation of the CAM programs to the Honeywell 2700 and the IBM 360-40. The work being done with the PDP-12 should result in yet new approaches to analyzing CAM data in small machines.

We are continuing to analyze important factors involved in implementing CAM in the school setting. As we learn more about this process, we are including relevant information in our developing manuals on the CAM system. Future manuals, for example, will include the details of monitoring systems designed for different purposes (e.g., the Jamesville-DeWitt and Greece models). Incidentally, the development of detailed engineering manuals is a normal part of the course of projects supported under New York State Experimental Programs.

These and other supporting activities will hopefully generate an increasing SED capability designed to modify evaluation procedures in the schools, through a series of demonstration projects. For example, when some of our software problems are solved, the BOCES's in Westchester and Nassau Counties may be able to demonstrate the feasibility of regional systems for monitoring the effectiveness of school programs at a particular level. Our efforts with the PDP-12 in Greece and the related projects at Ballston Spa and Jamesville-DeWitt potentially demonstrate the capability for bringing courses, programs and the learning activities of individuals under systematic control, without increased costs to the school. Hopefully, these efforts will find some application in designs

for the evaluation of special programs and instruction in the disadvantaged schools of the State, as well as in other special programs designed to ameliorate educational problems.

Campbell, D. T. Reforms as experiments. American Psychologist, 1969 (April), 24, 409-429.

Gorth, W. P., Dumont, F., & Wightman, L.E. Improving educational quality through Comprehensive Achievement Monitoring: a proposal for a statewide demonstration. Project CAM, Technical Memorandum TM-23. Amherst, Massachusetts: University of Massachusetts, November 1969.

Gorth, W. P., & Wightman, L.E. CAM described for state level evaluation of urban education projects. Project CAM, Technical Memorandum No. TM-20. Amherst, Massachusetts: University of Massachusetts, April 1969.

Gottman, J. M., McFall, R. M., & Barnett, J. T. Design and analysis of research using time series. Psychological Bulletin, 1969, 72 (4), 299-306.

New York State Education Department. Educational disadvantage in New York State: a two year report of the Pupil Evaluation Program test results. Albany, New York: Division of Educational Testing, Regents Examination and Scholarship Center, December 1968.

New York State Education Department. The educationally disadvantaged in New York State: the scope of the 1966 Pupil Evaluation Program test results. Albany, New York: Division of Educational Testing, Regents Examination and Scholarship Center, December 1967.

New York State Education Department. School administrator's manual: Part 2 - guide to interpretation of Pupil Evaluation Program test results. Albany, New York: Bureau of Pupil Testing and Advisory Services, The State Education Department, The University of the State of New York, January 1970.

New York State Education Department. Test results of the 1965 Pupil Evaluation Program in New York State: preliminary overview. Albany, New York: Division of Education Testing, Regents Examination and Scholarship Center, January 1967.

O'Reilly, R. P., Schriber, P. E., Gorth, W. P., & Wightman, L.E. Guide for implementing the Comprehensive Achievement Monitoring system. Draft copy. Albany, New York: Division of Research, The State Education Department, The University of the State of New York, November 1969.

Wilson, A. B. Educational consequences of segregation in a California community. In United States Commission on Civil Rights, Racial isolation in the public schools. Part 2. Washington, D. C.: U. S. Government Printing Office, 1967. Pp. 165-206.